
The Null Hypothesis

A clear explanation of an often misunderstood factor in statistical analysis.



Ludwig Krippahl is the president of the Portuguese Skeptics Association. He's a biochemistry PhD student and works in molecular modeling research.

Though often unappreciated, the null hypothesis has a crucial role in statistical analysis. I'll try to illustrate this with an example. Imagine we flip a coin ten times and get ten tails. This is suspicious, so we propose the following hypothesis:

H1- There is something wrong with this coin.

Now we need some way to evaluate our hypothesis. At first sight we would like a hypothesis that is likely to be true, and we could try to evaluate the probability of this hypothesis being true given the results we obtained.

But if we think about it it's evidently a bad idea. It makes no sense to consider the hypothesis as something that flips from true to false with a given probability depending on the observation, and furthermore we have no way to measure these probabilities.

Since we want the hypothesis to explain the observation, it's best to do it the other way around and ask, "If our hypothesis is true, how likely is this observation?" Now all we have to do is to calculate the probability of getting so many tails assuming that

there is something wrong with this coin.

The problem is that we can't. We could do this if we knew exactly what was wrong with the coin and how it affected the probability of heads or tails results, but we have no data on that so we can't calculate the effect.

However, if the coin were perfectly straight and balanced things would be easy. In this case the probability of heads would be the same as of tails, 1 in 2, and we could calculate everything. So let's consider two hypotheses instead of only one:

H1- There is something wrong with this coin.

H0- There is nothing wrong with this coin.

These two cover all the possibilities as far as the coin is concerned, and only one can be true. And though we can't measure directly how well our hypothesis (H1) explains the observations, we can measure how badly the other hypothesis (H0) does.

H0 is, of course, the null hypothesis.

If H0 is true the probability of heads is 1/2 per throw. For 10 throws we just have to multiply 1/2 by itself ten times, which is 1/1024, approximately 0.1%, and we can reject H0 with 99.9% confidence. As we saw above, this doesn't mean that H0 is 99.9% false, but that if H0 were true 99.9% of the results would be closer to the expected (half heads, half tails) than the results we got. Remember that we don't measure the probability of the hypothesis, but the probability of the results if the hypothesis were true.

In this case this probability is quite low, so we are justified in suspecting that there is something wrong with the coin. In short, the null hypothesis is that nothing special is happening, allowing us to calculate the likelihood of getting our results by chance. If these are too unlikely we can reject the null hypothesis and conclude that something is going on.

Limitations

Although very useful, this method has some important limitations. For one thing, rejecting the null hypothesis does not mean necessarily that our initial hypothesis is correct. In our coin example, we implicitly assumed that any problem was with the coin and not with the way it was tossed, how long it fell, where it fell, or other possibilities.

If the effect is strong it's easy to control the conditions so that other

factors can be ignored. Our coin example would work well even in practice and not just as a thought experiment. We can see that the effect is strong because we get very significant results even with a small sample of 10 throws.

ESP or Astrology studies, for example, suffer from this problem. Statistical significance in these cases comes only with very large data sets, indicating that the effect is very small. A single marked card in 52 increases the odds of guessing correctly by nearly 0.5%, and often the reported effects are much smaller than this. Such small effects are very hard to control, so the null hypothesis may easily be rejected because of a problem with the experiment.

Another thing to consider is that even unlikely results eventually crop up if we try long enough. If the null-hypothesis is true and is tested in 100 independent experiments we can expect 5 to reject it with 95% confidence and one of them with 99% confidence on chance alone.

The File Cabinet Effect

This would not be a problem if we knew about all 100 experiments. Since this outcome is also predicted by the null-hypothesis a global analysis would lead us to the right conclusion. However, the few experiments that rejected the null-hypothesis are more likely to be published than the many that didn't. This is called the "file cabinet effect",

as the uninteresting majority will end up stored somewhere and never be published. This bias in favour of positive results makes it necessary to have independent confirmation before we can conclude the null-hypothesis is wrong.

In conclusion, we should be suspicious of small effects that can only be detected in very large samples. Even if the null-hypothesis can be confidently rejected, this may only reflect a small problem with the experiment.

We should also look for independent confirmation. After all, a published positive result doesn't tell us how many negative results are left in the file cabinet, and a confirmatory experiment is interesting and likely to be published whether it confirms or refutes the initial results.

Finally, we should be especially wary of meta-analyses, where surveys of published results are used to reject the null-hypothesis. Without access to the multitude of file cabinets where all the negative results are hidden, the sample is too biased for a reliable conclusion.

Note: I made no distinction between two-tailed and one-tailed tests. This is important in practice because it affects the probability values, but it's not necessary to understand the process.

